Projet Statistiques DU-SDA

Rola Monzer

31 janvier 2022

Introduction

L'objectif de cet exercice est l'analyse du poids des nouveaux-nés selon différentes variables explicatives comme le nombre de semaine de gestation, l'âge du père, l'âge de la mère, fumeuse ou non...

1 Analyse Descriptive

Analyse descriptive des données graphiquement et statistiquement.

1.1 Réaliser une analyse descriptive approfondie des variables

Il existe 16 variables dans notre dataset où aucune d'entre elles n'a une valeur NAN.
La première variable est ID qui représente le numéro du nouveau-né, cette variable nous intéresse pas comme l'identité du nouveau-né est fixer aussi par l'index du data-frame. La deuxième variable représente notre target Birthweight ou bien le poids du nouveau-né. Les variables 3 et 4 représentent la circonférence de la tête du nouveau-né et le nombre des semaines de gestation. Tous ces variables correspond au informations du nouveau-né leurs descriptions statistiques est illustré dans la figure 1.

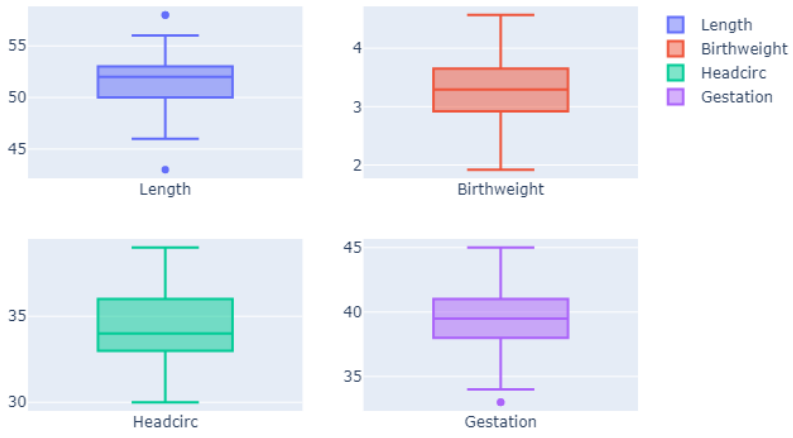


FIGURE 1 – Boxplots des différents variables qui contient les informations des nouveau-nés.

Ensuite, les autres variables quantitatives nous donne des informations sur les parents : l'âge, la taille et le nombre de cigarettes consommées par jour pour chacun des parents. En plus, on a le poids de la mère avant la grossesse et le nombre d'années d'éducations du père. Les descriptions statistiques de ces variables sont illustrées dans la figure 2.

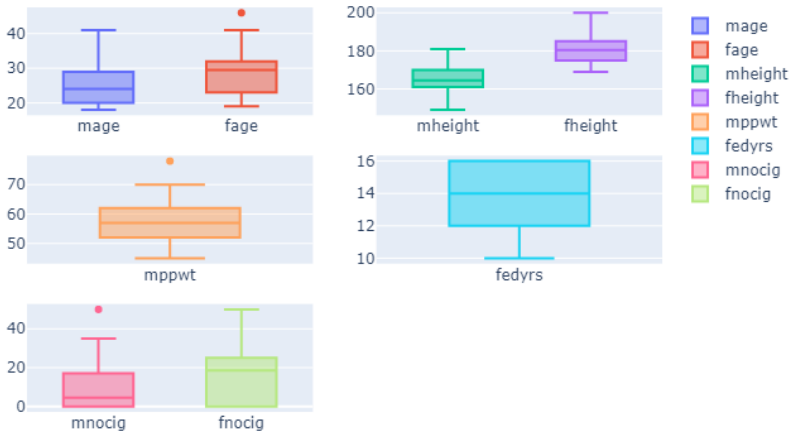


FIGURE 2 – Boxplots des différents variables qui contient les informations des parents.

Il nous reste 3 variables binaire à inspecté : smoker est True si la mère fume, lowbwt est True si le poids du nouveau-né est considérer faible et mage35 est True si l'âge de la mère est égale ou dépasse 35. Les histogrammes de la figure 3 montrent leurs distributions.

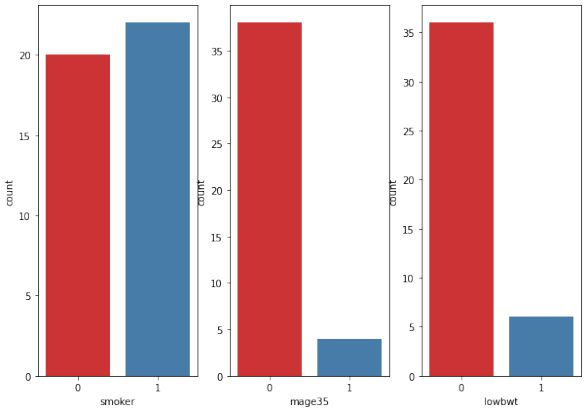


FIGURE 3 – Histogrammes des variables binaire.

1.2 Faire une analyse par rapport à la variable à expliquer "birthweight"

On connaît déjà de la figure 1 que les valeurs de la variable Birthweight sont entre 1.92 et 4.57 kg avec 3.29 kg comme médiane. On calcule la moyenne qui est égale à 3.31 kg. Ensuite, on a choisi de visualiser les changements de la variable Birthweight en fonction des autres variables (sauf ID) dans la figure 4.

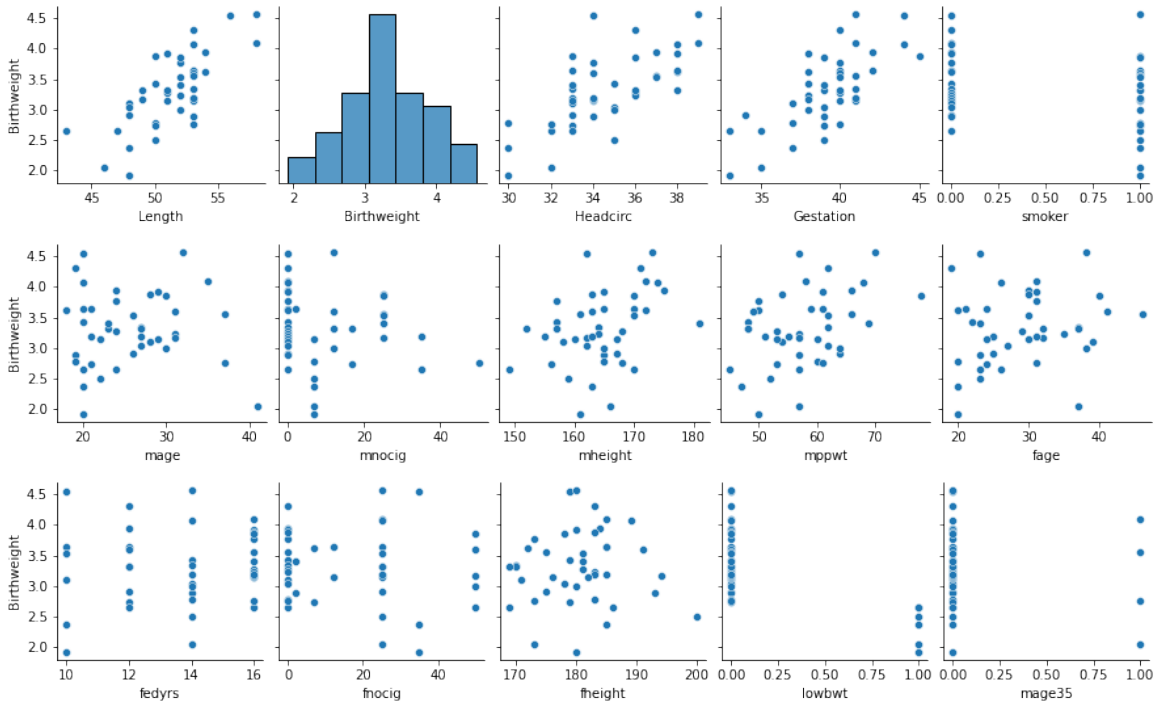


FIGURE 4 – Paiplot illustrant la dépendance de Birthweight en fonction des différents variables

1.3 Faire une analyse de corrélation des variables

On utilise la fonction corr() du pandas pour calculer la corrélation entre tous les variables. Cette fonction est très importante pour l'implémentation de notre modèle comme on peut trouver les variables qui affectent les plus notre target. La figure 5 illustre les résultats de la fonction de corrélation.

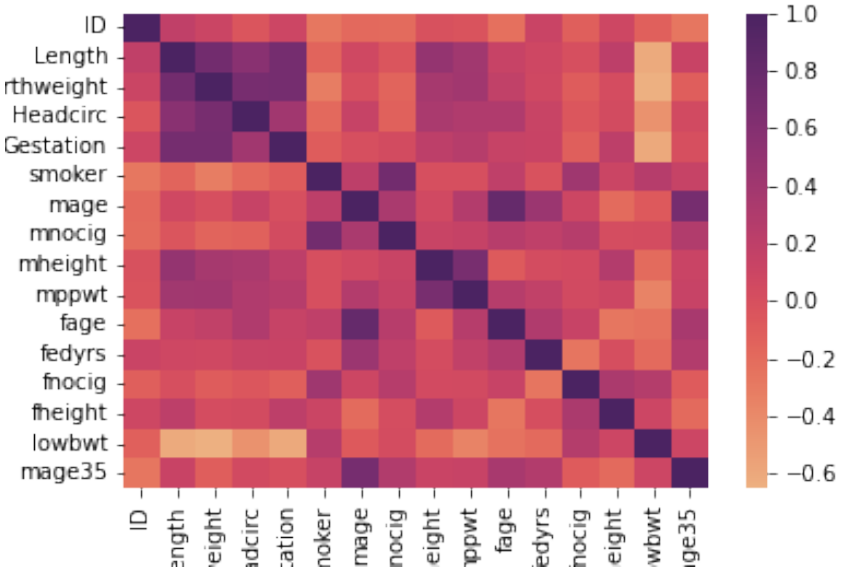


FIGURE 5 – Heatmap représentant les corrélations entre tous les variables.

2 Features Engineering

Définir les variables catégorielles.
En effet, tous nos variables sont soit quantitative soit binaire mais c'est toujours possible de catégoriser les variables quantitative.

2.1 Modifier les variables qualitatives nominales en variables catégorielles

Ici, on a choisi de catégoriser des variables qui ne sont pas très corréler avec notre target le poids des nouveaux-nés tel que l'âge (coefficient de corrélation <0.2), la taille du père, le nombre de cigarettes consommées par jour par chacun des parents, et le nombre d'années d'éducations du père.

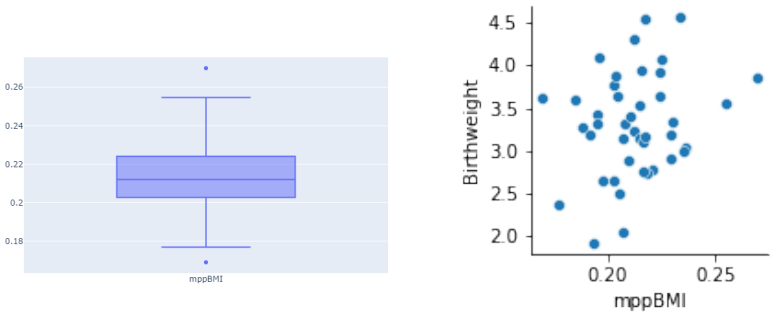
On a choisi d'avoir 2 catégories pour le nombre d'années d'éducations du père (voir la section 2.2), 3 catégories pour l'âge et le nombre de cigarettes consommées par jour par chacun des parents, et 4 catégories pour la taille du père. On note qu'on a supprimer la variable ID de notre dataframe.

2.2 Modifier les variables catégorielles en variables binaires (si nécessaire)

En effet, la variable fedyrs ne semble pas très intéressante et c'est pas trop corréler a notre target (coefficient de corrélation = 0.07) alors on choisi de remplacer cette variable par une variable binaire fhighed qui serait True si le nombre d'années d'éducation du père est supérieur ou égale à 14.

2.3 Créer de nouvelles variables pertinentes pour l'analyse

Une des variables intuitive lorsque l'on dispose des valeurs du poids et de la taille c'est l'indice de masse corporelle ou IMC. Dans notre dataset c'est possible de calculer l'IMC du nouveau-né et celui de la mère avant la grossesse. En effet, pour calculer l'IMC du nouveau-né on a besoin de son poids qui est la variable target, alors ça serait pas efficace de l'utiliser afin de prédire notre traget a cause de la colinéarité. Mais l'IMC de la mère avant la grossesse peut être intéressant et alors on crée la variable mppBMI .
L'analyse statistique de la variable mppBMI est représenter par la figure 6a, et sa corrélation avec Birthweight par la figure 6b.



(a) Boxplot de la variable mppBMI (b) Birthweight en fonction de mppBMI

FIGURE 6 – Graphiques explicatives de la variable mppBMI

3 Test de Student

Définir l'hypothèse nulle et mesurer la significativité du test.
L'hypothèse nulle est que le poids du nouveau-né est indépendant de si la mère est fumeuse.

3.1 Comparer le groupe des mères fumeuses et des mères non-fumeuses sur le poids des nouveaux-nés, que remarquez-vous ?

On choisi de faire un two-sample t-test, comme il s'agit de deux groupes : fumeuse et non-fumeuse, et notre objectif est de comparer leurs moyennes. On note que pour tout les tests on considère 5% comme la p-value minimal pour accepter l'hypothèse nulle.

Premièrement, On compare les variance des deux groupes qui sont pareil à 0.12 près. Après, on utilise la fonction ttest_ind de scipy pour obtenir la p-value qui est égale à 4,26% alors on rejette l'hypothèse nulle et donc le poids des nouveaux-nés dépend de si la mère fume ou pas. En fait, comme le test statistique est négatif = -2.09, les mères fumeuses ont tendance à avoir des nouveau-nés ayant un poids faible.

4 Test Chi-Deux

Définir l'hypothèse nulle et mesurer la significativité du test.
L'hypothèse nulle est que les variables lowbwt et smoker sont indépendant.

4.1 Réaliser un test de chi-deux pour mesurer l'indépendance des variables entre fumeuses et le poids léger des nouveaux-nés (lowbwt)

Les deux variables lowbwt et smoker sont des variables binaires donc le test chi-deux est le test approprier, et on n'a pas besoin de les transformer en variables catégorielles.
On utilise la fonction crosstab de pandas pour avoir un tableau de fréquences d'occurrence des deux variables en question, on illustre ce tableau dans la figure 7. Ensuite, on utilise la fonction chi2_contingency de scipy pour calculer la p-value qui est égale a 23.08% alors on accepte l'hypothèse nulle que les variables lowbwt et smoker sont indépendantes.

On note que dans ce cas on risque toujours avoir une erreur de type I, c'est à dire qu'on accepte à tord l'hypothèse nulle. Nous soupçonnons que c'est le cas comme la variable lowbwt dépend uniquement de Birthweight, et d'après la section 3 les variables smoker et birthweight ne sont pas indépendantes.

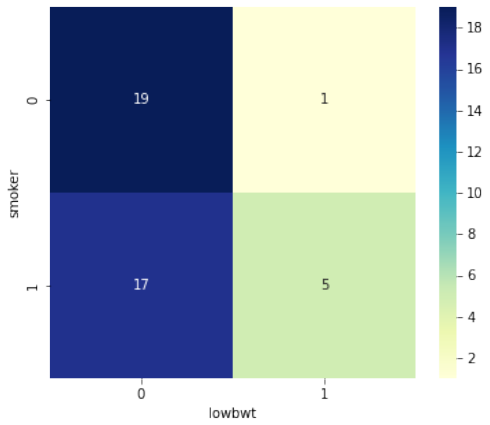


FIGURE 7 – Tableau représentant les fréquences d'occurrence des deux variables.

5 Test de normalité

Définir l'hypothèse nulle et mesurer la significativité du test.
L'hypothèse nulle est que la variable Birthweight suit une loi normale.

5.1 Définir si birthweight suit une loi normale

Pour le tester on exécute le teste de normalité de Shapiro-Wilk où on trouve p-value = 96% . Alors, on accepte l'hypothèse nulle et donc Birthweight suit une loi normale.

En plus, dans la figure 8 on compare l'histogramme du poids des nouveaux-nés avec la courbe d'une distribution normale ayant la même moyenne et le même écart type. On note que l'histogramme ressemble a une distribution normale ce qui confirme que l'hypothèse nulle est valide.

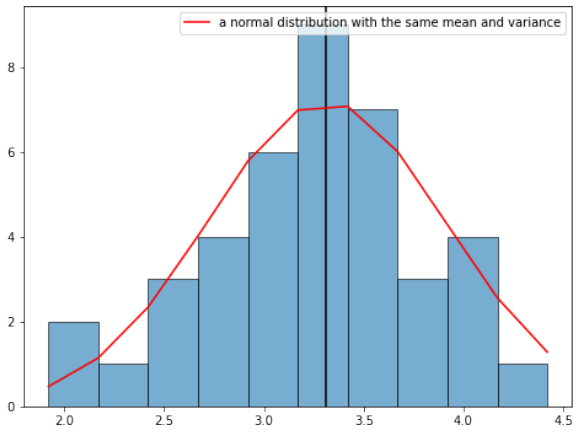


FIGURE 8 – Histogramme représentant la distribution de la variable Birthweight et la distribution normale qui lui correspond.

6 Régression linéaire

Entraîner un modèle de régression linéaire et définir le choix et la significativité des variables explicatives. Une évaluation de votre modèle sera nécessaire.

6.1 Vérifier que les hypothèses sont vérifiées

On définit chaque hypothèse nulle par l'indépendance entre la variable explicative approprier et notre target Birthweight.

6.2 Implémenter une régression linéaire

On choisi d'utiliser la librairie statsmodels pour implémenter notre modèle. On a opté pour une méthode d'essai et d'erreur afin d'optimiser notre modèle.

On note qu'on utilisera pas la variable lowbwt comme c'est uniquement dépendante de notre target et alors on aura des problème de colinéarité. En plus, si on a la valeur de cette variable ça veux dire qu'on connais déjà le poids du nouveau-né.

On commence par considérer que les variables Length, Headcirc et Gestation comme les variables explicatifs. On obtient la valeur 0.713 de r-squared et on trouve que la variable Length a une p-value significatif tandis que Headcirc et Gestation ont des faibles p-values. Alors on accepte l'hypothèse nulle pour les deux dernier mais on la rejette pour la variable Length .

On fait un autre essai avec la variable smoker et tous les variables qu'on a créer dan la section 2 : mppBMI , magegrp , fagegrp ... Ces variables remplace les autres variables initiales qui contenait les informations des parents. On obtient la valeur 0.211 de r-squared, qui signifie que notre modèle est moins optimiser. En plus, on trouve que tous les p-values sont supérieurs à 5% et alors on rejette l'hypothèse nulle pour tous les variables. Donc tous les variables catégorielles (plus la variable mppBMI) et la variable "Birthweight" sont dépendent.

Ensuite, on utilise tous les variables précédents sans exclure les variables Headcirc et Gestation même si ils sont à priori indépendant de notre target. On aura un modèle ayant r-squared = 0.804 avec toujours des p-values faibles pour Headcirc et Gestation .

Notre dernière tentative sera de supprimer ces deux derniers variables de notre modèle. On obtient la valeur 0.657 pour r-squared, qui est plus faible que celle d'avant, donc en supprimant les variables Headcirc et Gestation notre modèle sera moins optimiser. En plus, la p-value associée avec Length n'est plus significative, et si on supprime cette variable on revient à notre deuxième tentative qui est la moins optimiser.

Par conséquent, on choisi la troisième tentative comme notre modèle de régression linéaire ayant le plus de crédibilité pour prédire le poids d'un nouveau-né.

6.3 Vérifier la significativité des variables explicatives

Notre modèle dépend de plusieurs variables : Length, Headcirc, Gestation, smoker, magegrp, fagegrp, mnociggrp, fnociggrp, fheightgrp, fhighed, mppBMI.

On a déjà parler des deux variables Headcirc et Gestation ayant des faibles p-values et alors qui sont pas trop significatifs. Mais, quand on compare les autres p-valueson trouve que les variables fagegrp, fnociggrp et fheightgrp ont les p-values les plus haut, mais des coefficient qui sont faible. Donc, notre target dépend de ces variables faiblement. On note que plus p-value est élevé, plus on est sûrs de rejeter l'hypothèse nulle.

D'autre part la variable mppBMI a une valeur significatif pour p-value et pour le coefficient, cette variable est suivi par les variable smoker, magegrp et mnociggrp qui sont tous des variables contenant les informations de la mère. Alors on peut conclure que les données maternelles ainsi que la variable Length sont les plus important pour prédire le poids d'un nouveau-né.

6.4 Évaluer votre modèle

Finalement, on exécute une cross-validation par implémentions d'une régression linaire en utilisant la librairie sklearn. On définit les features columns qui sont les mêmes variables utiliser pour implémenter notre modèle dans la partie précédente et on les fit dans le cadre du modèle LinearRegression de sklearn. On trouve les mêmes coefficients et la même valeur de r-squared. Donc, on peut confirmer que notre modèle est optimiser.

Conclusion

Dans cet exercice, on a analyser les différents variables du dataset « Birthweight.csv », et on a calculer et visualiser leurs corrélation avec notre target. En plus, on a catégoriser quelques variables et en créer des autres afin de mieux prédire le poids des nouveaux-nés. On a exécuter plusieurs tests statistiques avec différents hypothèses pour mieux comprendre si le faite que la mère fume affect le poids du nouveau-né. Et on a confirmer que le poids des nouveau-nés suit une loi normale. Enfin, on a implémenter un modèle de régression linéaire qui utilise les variables : Length, Headcirc, Gestation, smoker, magegrp, fagegrp, mnociggrp, fnociggrp, fheightgrp, fhighed et mppBMI pour prédire le poids du nouveau-né.